IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Richard K. Lincoln, Jr., PMP, MBA
January 5, 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data was gathered from the SpaceX website and a Wikipedia page with tables related to the Falcon 9 rocket. The data was cleaned, standardized, and analyzed. The data was randomly split into training and testing samples. Four models were used to determine the most accurate predictive model ensuring parameters for each model were tuned.
- Summary of all results
  - All models performed well with all models performing consistently with the exception of the K N-Neighbors model which scored slightly lower than the others. Thus, it is possible to predict a successful booster landing with reasonable (82%) accuracy knowing the Launch Site, Payload, and Booster Version.

Introduction

- Project background and context
  - The cost of putting a payload into orbit is impacted by the success or failure of the booster being able to land successfully and be reused. Thus, if we can predict the probability of a successful landing, we can more fully understand the probable cost of a launch.
- Problems you want to find answers to
  - What is the probability of a successful landing?
  - Is the probability of a successful landing effected by
    - Launch Site
    - Payload
    - Flight Number
    - Orbit

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected using the SpaceX REST API
  - Launch Data including Booster, Payload, Launch Site, and Landing Success
- Data was collected by scaping Wikipedia
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Historical Launch Records of the Falcon 9

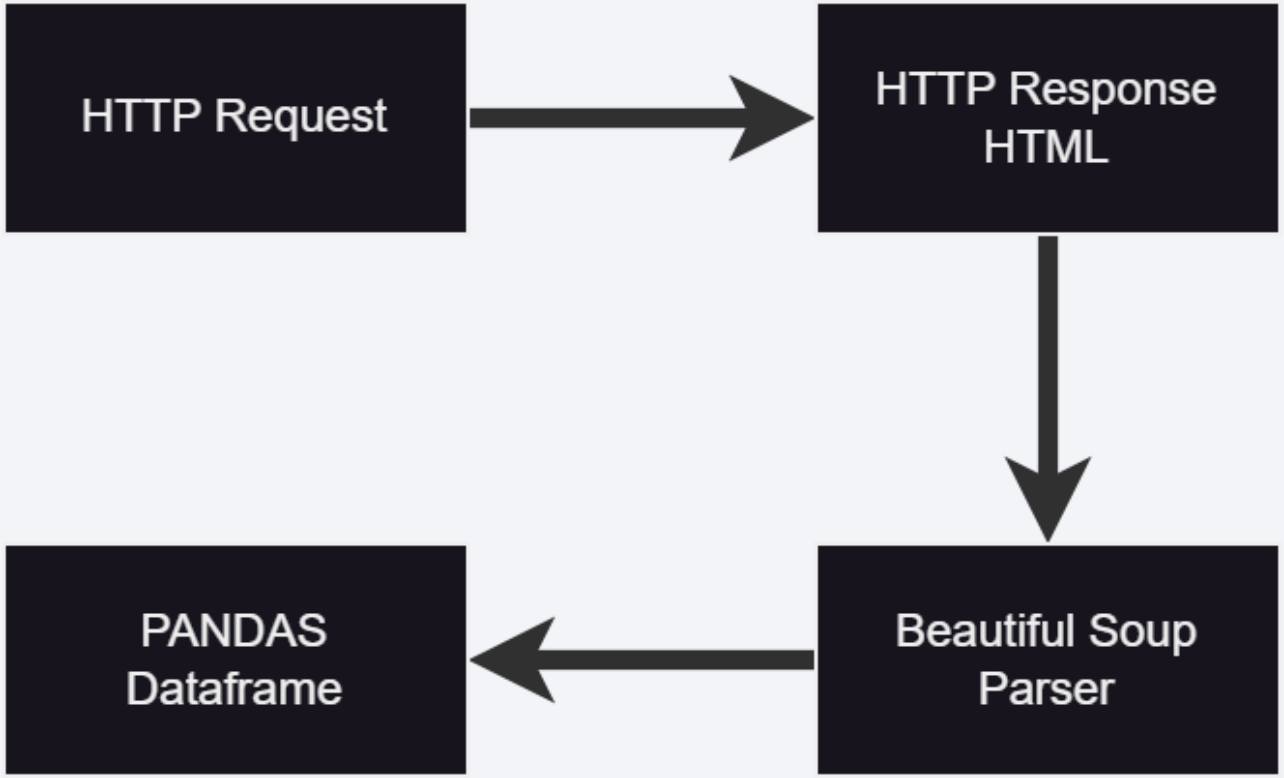
Data Collection – SpaceX API

- Data access can be simplified to three steps as seen in the simplified process flow.

- The Jupiter Notebook may be accessed here
  - (https://github.com/rkljr/DS-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb)

Data Collection - Scraping

- Data is gathered via a simplified four step process

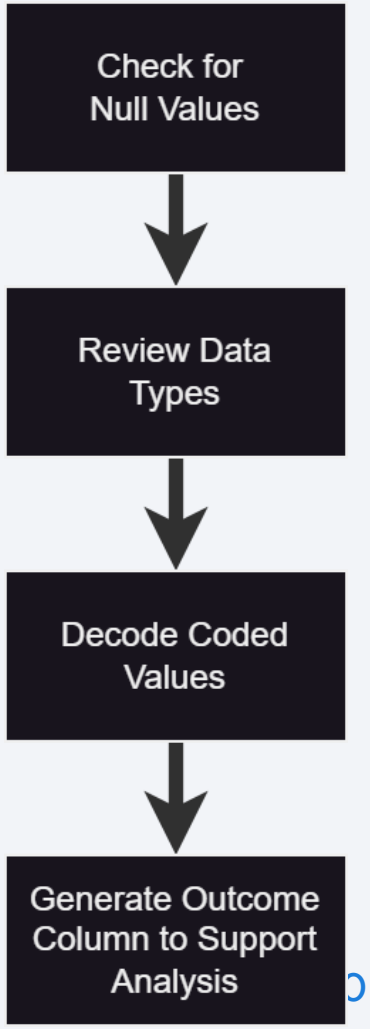


- The Jupiter notebook can be found here.
- https://github.com/rkljr/DS-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Once data were in a dataframe, the completeness of the data was checked and null values filled in or dropped as needed. Coded columns were decoded into meaninful values. Most importantly, launch outcomes were categorized into successs and failures in a new column to be used as the dependent variable in future analysis. .

- The Jupiter notebook can be found here.
  - https://github.com/rkljr/DS-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Numerous Scatter plots were used to visualize launches and understand success rates in terms of Launch Sites, Payloads, Orbit, and Flight Number of Launches (a proxy for experience).
- Scatter plots were also use to evaulate relationships between Orbit and Payload type and Orbit and Flight Number.
- A line plot was generated to visualize the trend in annual success rate over time.
- The Jupiter notebook can be found here.
    - https://github.com/rkljr/DS-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- SQL Queries
  - Distinct launch sites (SELECT DISTINCT...)
  - Launch sites beginning with 'CCA' (WHERE "Launch Site" LIKE 'CCA%')
  - Total payload mass of all launches for NASA (CRS) (SUM and WHERE "Customer" = ...)
  - Average payload mass carried by the F9 v1.1 booster (AVG and WHERE "Booster"= ...)
  - The first sucessful landing date (MIN and WHERE "Landing_Outcome" = ...)
  - Listing of boosters which resulted in successful drone boat landings (WHERE "Booster" = '" AND "Payload" >..))
  - The total number of landings by Landing_Outcome (GROUP BY "Landing_Outcome")
  - The names of the boosters which have carried the maximum payload (DISTINCT, Subquery)
  - Listing of failed landings in 2015 on drone ships by month (substr, WHERE
  - Rank listing of outcomes between two dates (ORDER BY, BETWEEN)
- The Jupitr notebook can be found here.
  - https://github.com/rkljr/DS-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

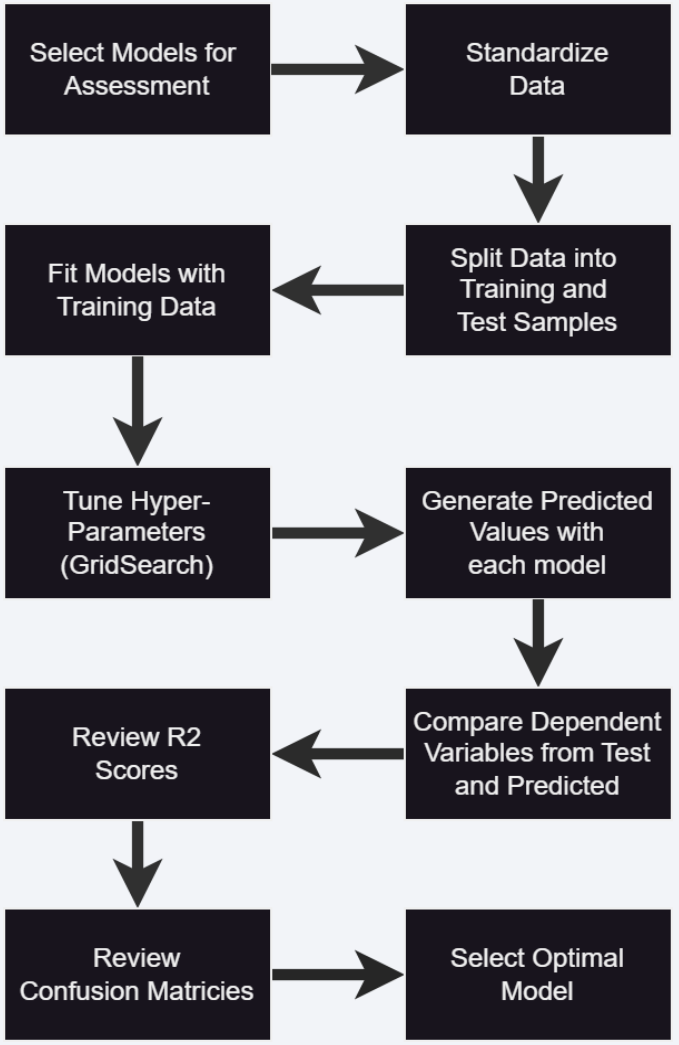
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- An overall view of the geographic location of launch sites using markers to indicate the location of each launch site in order to see the geographic dispersion of launch sites.
- Added clusters to permit the display of landing outcomes for each launch site. Using clusters permits the documentation of multiple values at a single geographic point.
- Created a function to display coordiates in latitude and lontitude of the mouse location to assist in capturing the location to be used in creating a line measuring the distance between to features.
- The Juiper notebook can be found here.
  - https://github.com/rkljr/DS-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Provided the user with a dropdown menu to select All or a specific Launch Site. This permits the user to review success rates by site or in aggrate.
  - Success rates are displayed in a pie chart to inform the user in a visual way the rate (%) of success and failures. (see pie chart here)
- Provded the user with a slider to permit them to select various ranges of payloads. This permits the user to review success rates by payload for any desired range of payloads
  - The payload succes status is displayed in a scatter plot making it visually easy for the user to see which payloads and boosters were successful. (see scatter plot here)
- Explain why you added those plots and interactions
- The python script can be found here.
  - https://github.com/rkljr/DS-Capstone/blob/main/spacex_dash_app.py

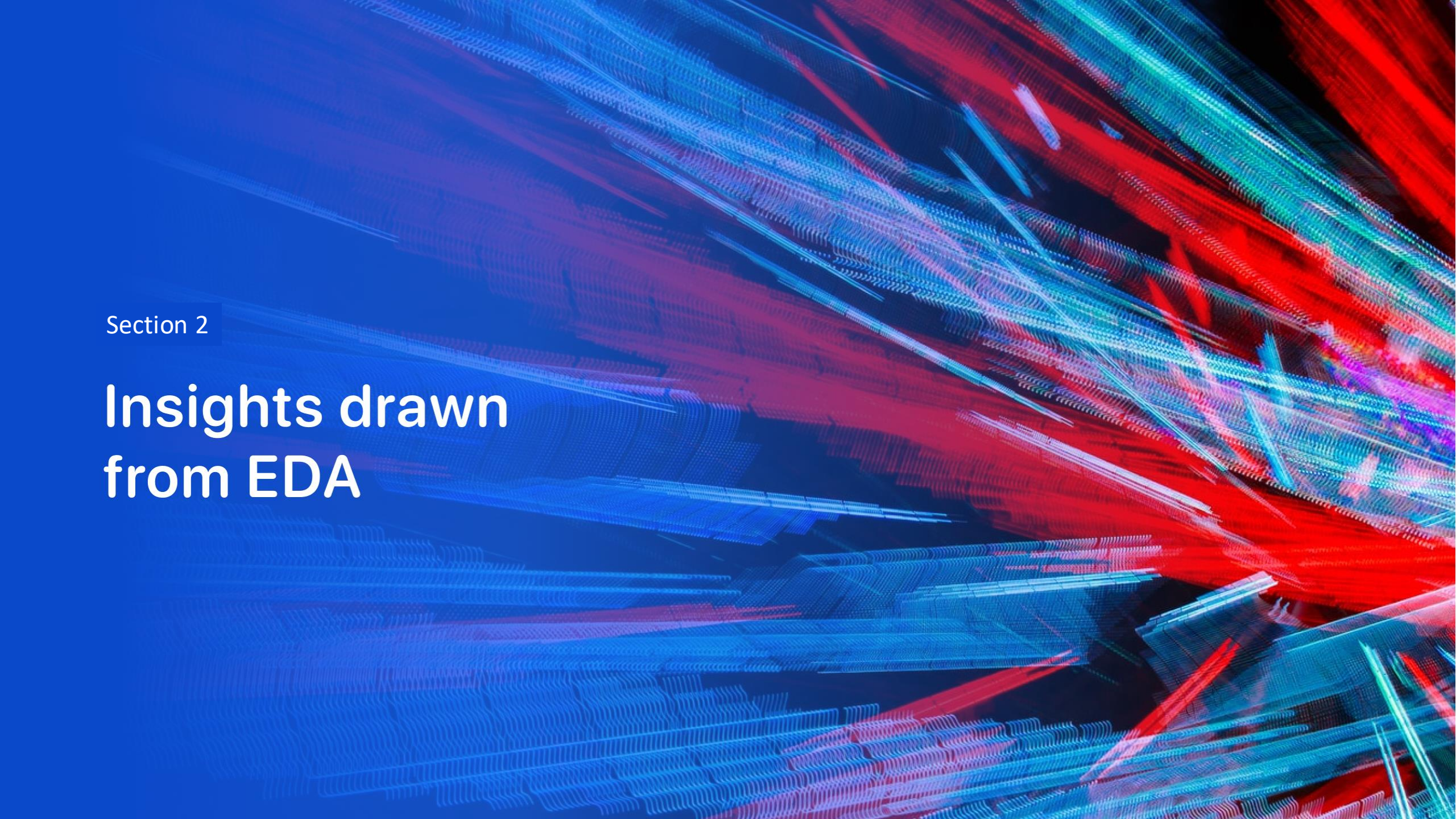
Predictive Analysis (Classification)

- The following models were evaulated and compared using the flow dipicted in the flow chart.
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K N-Neighbors
- All models performed well with K N-Neighbors being the least accurate (0.61)
- The Jupiter notebook can be found here.
  - https://github.com/rkljr/DS-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

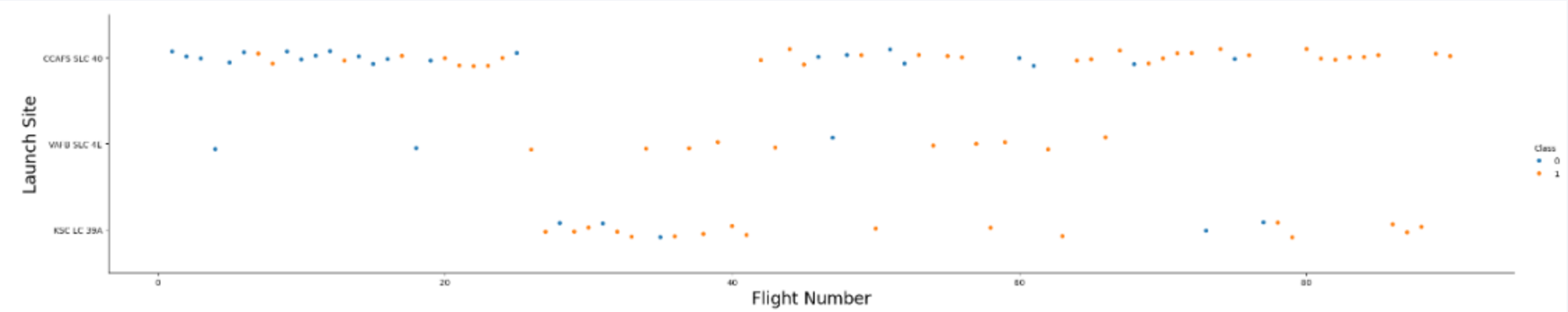


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

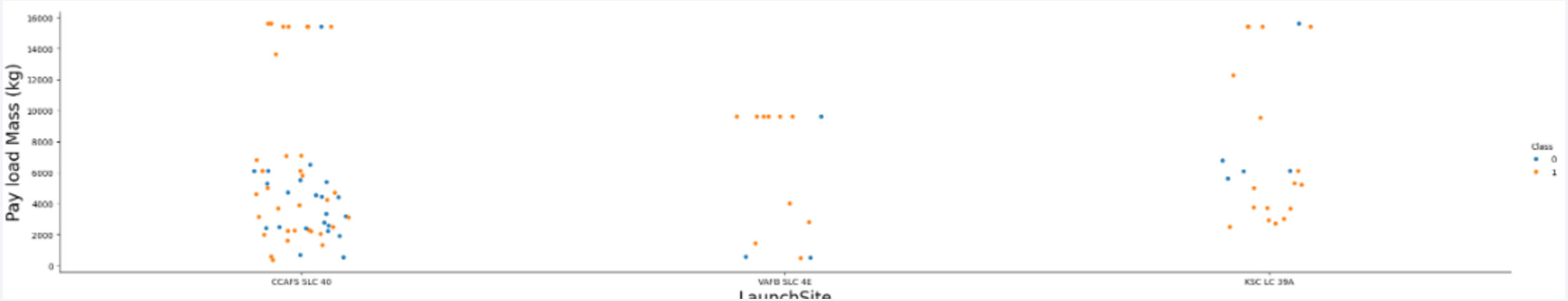

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The Y axis is the Launch Site and the X Axis is the Flight Number of the number of cumulative launches from the site. Points are colored green (landing failure) or amber (landing success).
- The cumulative number of flights from a Launch Site appears to be a predictor of landing success.

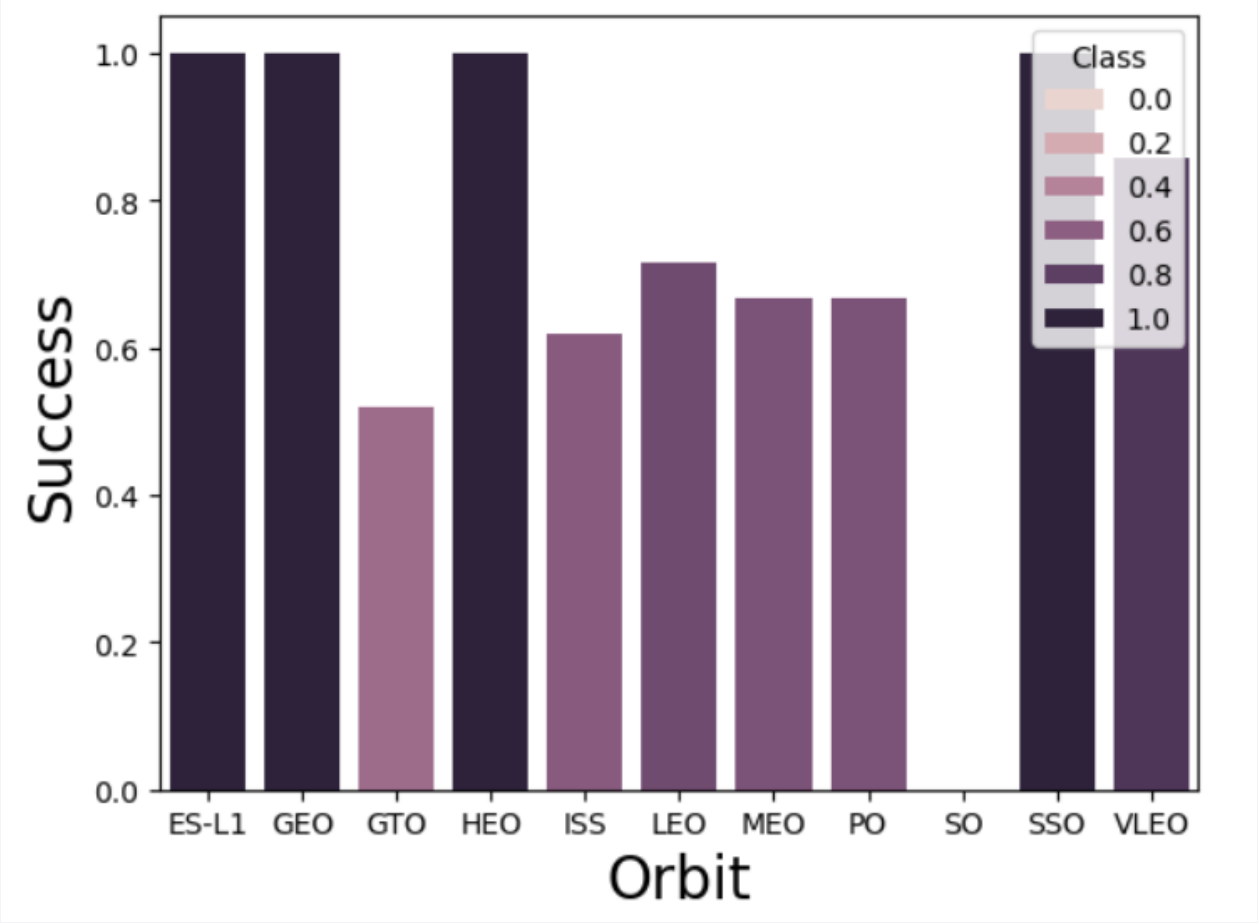
Payload vs. Launch Site



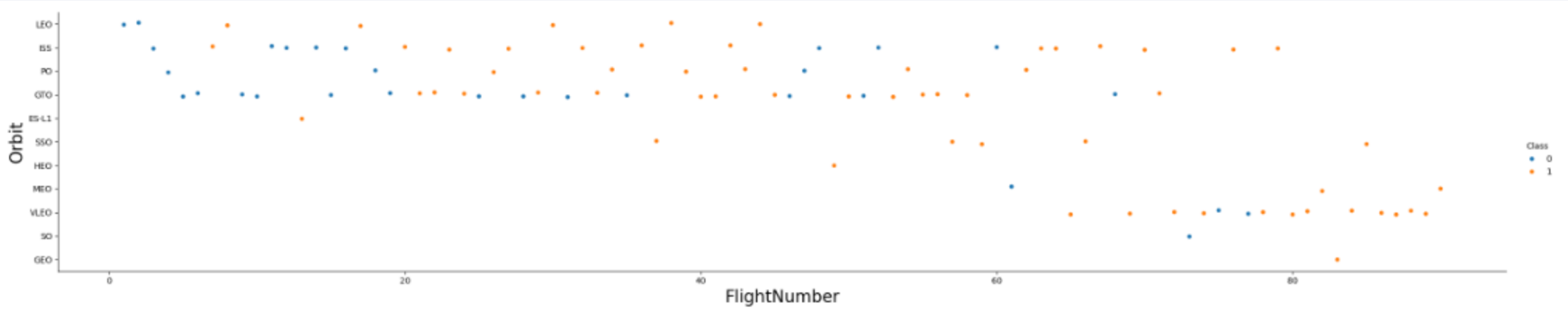
- The X axis is the Launch Site and the Y Axis is the Payload Mass (kg) carried by the rocket. Points are colored green (landing failure) or amber (landing success).
- It appears that from some Launch Sites a higher payload may predict a landing success.

Success Rate vs. Orbit Type

- The Y axis plots the mean success of each orbit on the X axis.
- The Orbits ES-L1, GEO, HEO, and SSO have almost perfect landing success rates.

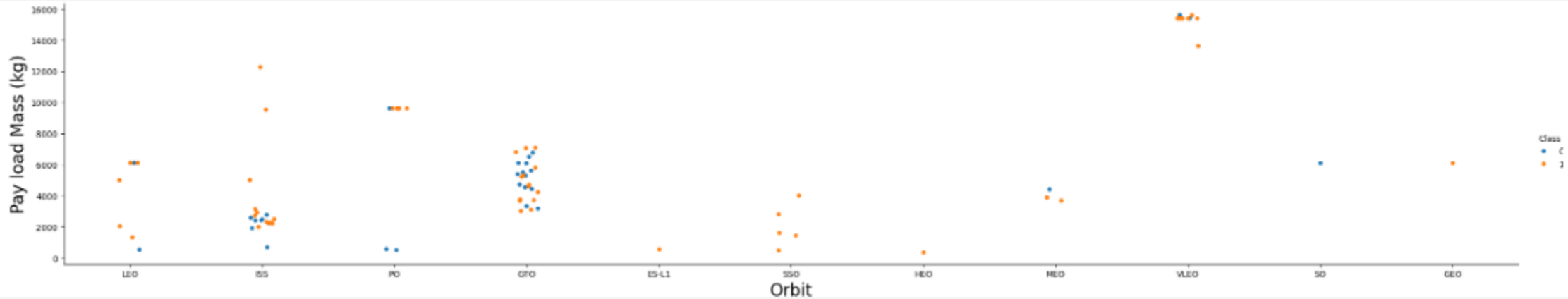


Flight Number vs. Orbit Type



- The Orbit is shown on the Y axis and the cumulative number of flights is shown on the X axis. The points are colored green (landing failure) and amber (landing success).
- The number of cumulative fligths appears to be a reasonable predictor of landing success.

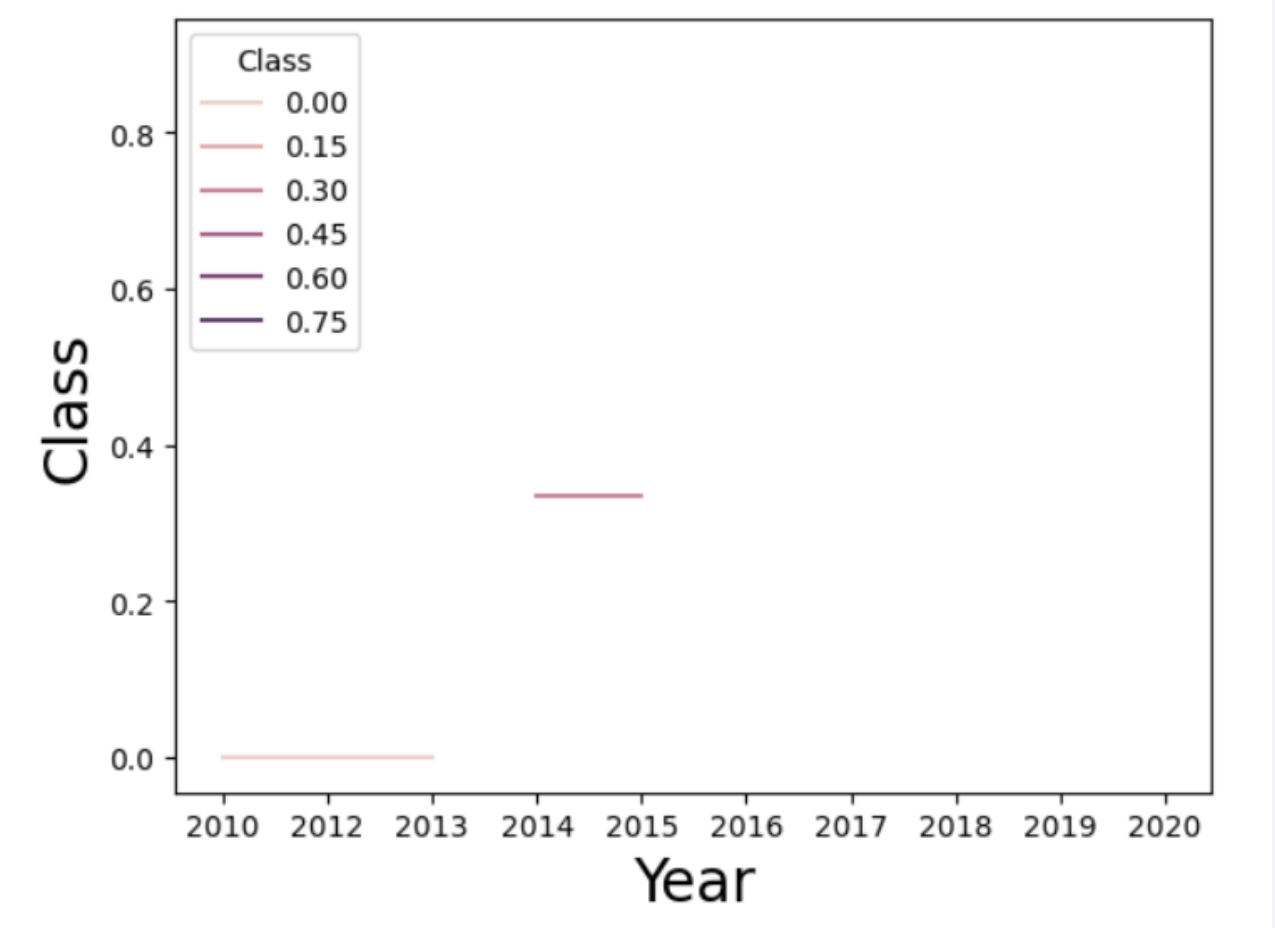
Payload vs. Orbit Type



- The Payload Mass (kg) is displayed on the Y axis and the Orbit type is displayed on the X axis. The landing success is represented by the point color (amber = success, green = failure).
- It is hard to draw a conslusion from this chart as there are too few points for some orbits. There may be some relationship between higher payloads and a higher landing success rate. This relationship should be futher assessed.

Launch Success Yearly Trend

- The annual mean landing success rate (Class) is presented on the Y axis and the year is presented on the X axis.
- Landing success rates appear to be correlated with year with success rates improving over time.



All Launch Site Names

- The query selects the distinct/unique Launch Site names from the SPACEXTABLE.

```
%sql SELECT  DISTINCT("Launch_Site") FROM SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- SELECT "Launch_Site" FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
- Select values from Launch Site in the SPACEXTABLE where the Launch Site begins with the characters "CCA". Only return five rows.

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

SUM("PAYLOAD_MASS__KG_")
45596

- Sum all Payload Mass values from the SPACEXTABLE where the customer is 'NASA (CRS)'

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

AVG("PAYLOAD_MASS__KG_")
2928.4

- Calculate the average of the Payload Mass in the SPACEXTABLE where the Booster Version is "F9 v1.1".

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

MIN("Date")
2015-12-22

- Retrieve the minimum (first) date from the SPACEXTABLE where the Landing Outcome was success.

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000
- Return the Booster Version field from the SPACEXTABLE where the booster landed on a drone ship successfully and the payload was greater than 4000kg and less than 6000kg.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- SELECT "Landing_Outcome", COUNT(*) FROM SPACEXTABLE GROUP BY "Landing_Outcome"
- Count the number of each Landing Outcome in the SPACEXTABLE

Landing_Outcome	COUNT(*)
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- SELECT DISTINCT("Booster_Version") FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE)
- Return a unique list of Booster Versions from the SPACEXTABLE there the Payload Mass carried is equal to the maximum payload mass in the SPACEXTABLE.

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

Month	"Landing_Outcomes"	"Booster_Versions"	Launch_Site
01	Landing_Outcomes	Booster_Versions	CCAFS LC-40
04	Landing_Outcomes	Booster_Versions	CCAFS LC-40

- SELECT substr(Date, 6,2) AS 'Month', "Landing_Outcomes", "Booster_Versions", "Launch_Site" FROM SPACEXTABLE WHERE substr(Date,0,5) = '2015' AND "Landing_Outcome" = 'Failure (drone ship)'
- Use the substr (sub-string) function to extract the four digit month from the string date and name the resultant field 'Month', and return the Landing Outcomes, Booster Version, and Launch Site fields from the SPACEXTABLE. Only return records from the year 2015 (year extracted from the date string using substr functon) and where Landing Outcome was 'Failure (drone ship)'.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- SELECT "Landing_Outcome", COUNT(*) FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY 2 DESC
- Return a count of the Landing Outcome from the SPACEXTABLE where the landing was between June 4, 2010 and March 20, 2017. Sort the values in decenting order based upon the counts.

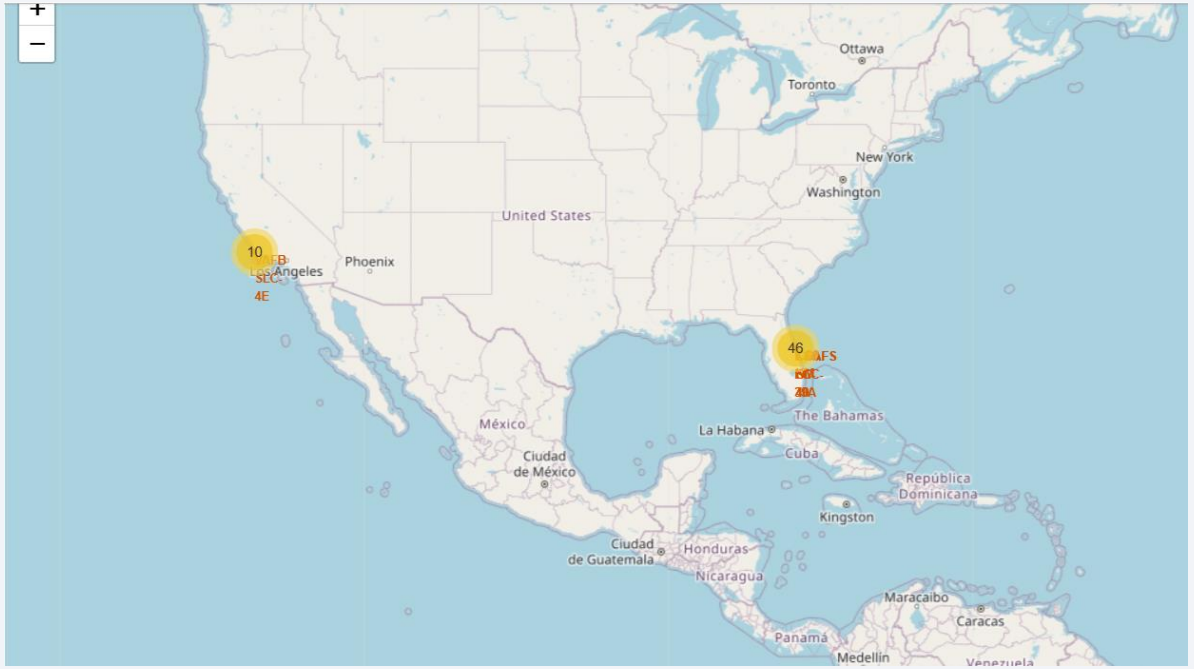
Landing_Outcome	COUNT(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

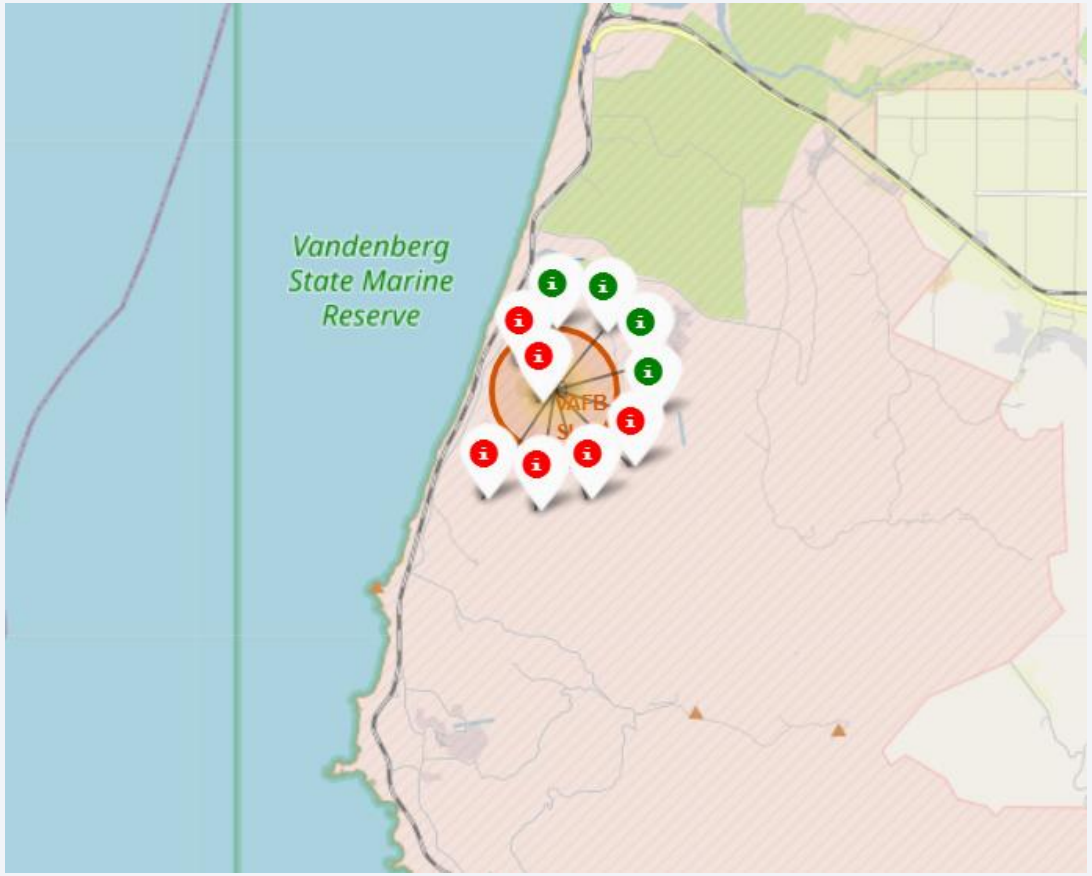
Launch Site Overview

- Launch sites are located along the west and east coasts
- There have been ten launches on the west coast and 46 on the east coast.



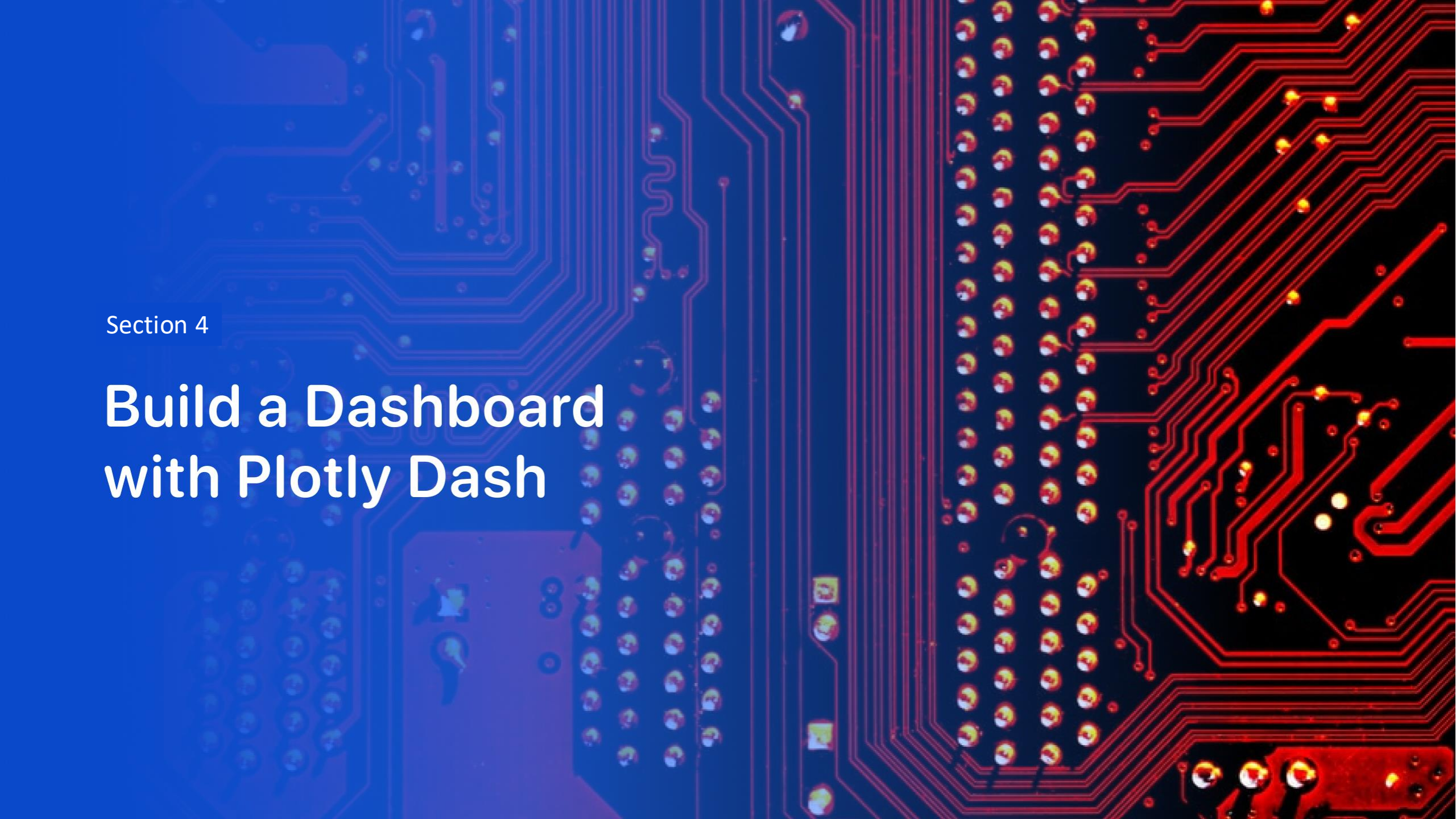
West Coast Launches

- There are a total of six launches with unsuccessful landings
- There are a total of four launches with successful landings

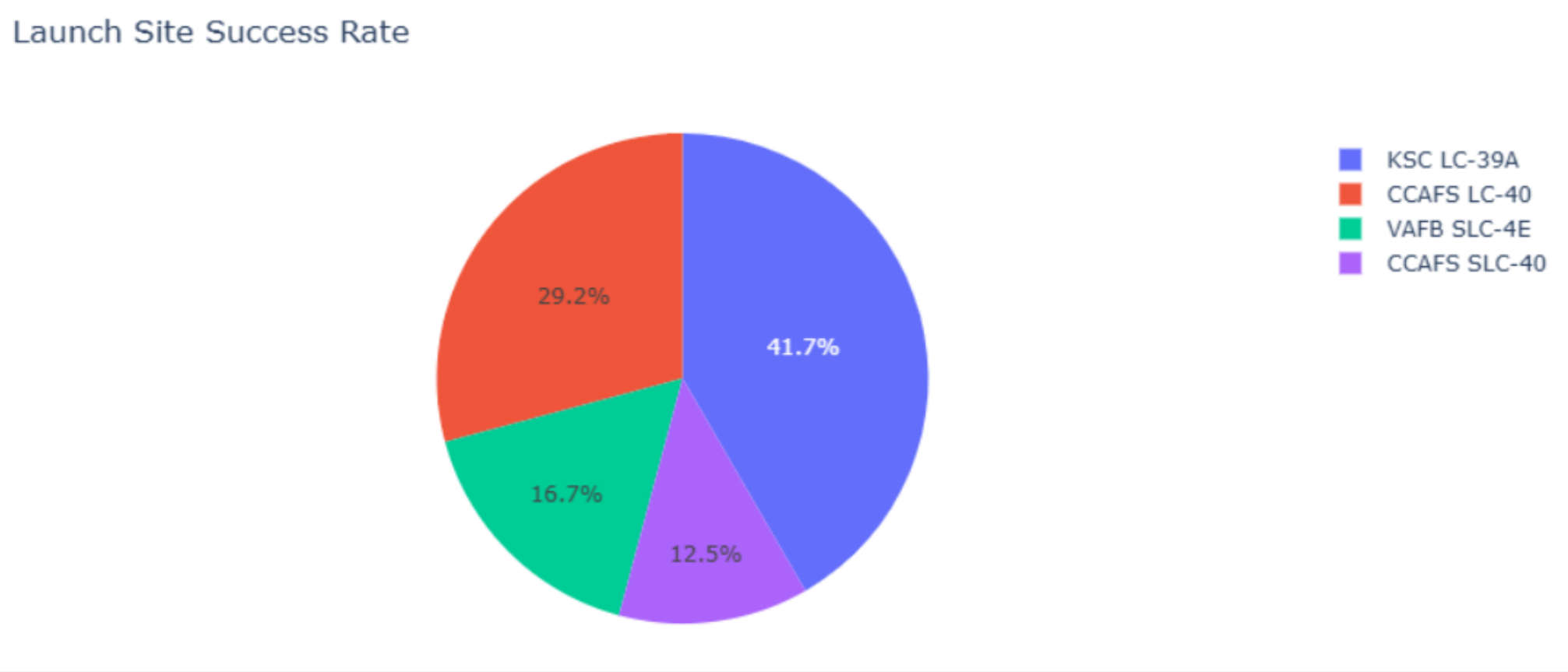


Launch Site KSC-LC 39A

- The nearest rail road is 0.73k from the launch site
- The nearest highway is 0.84k from the launch site.
- There coast is approximately 7.66k to the east of the launch (not pictured)
- There have been 13 launches from this site

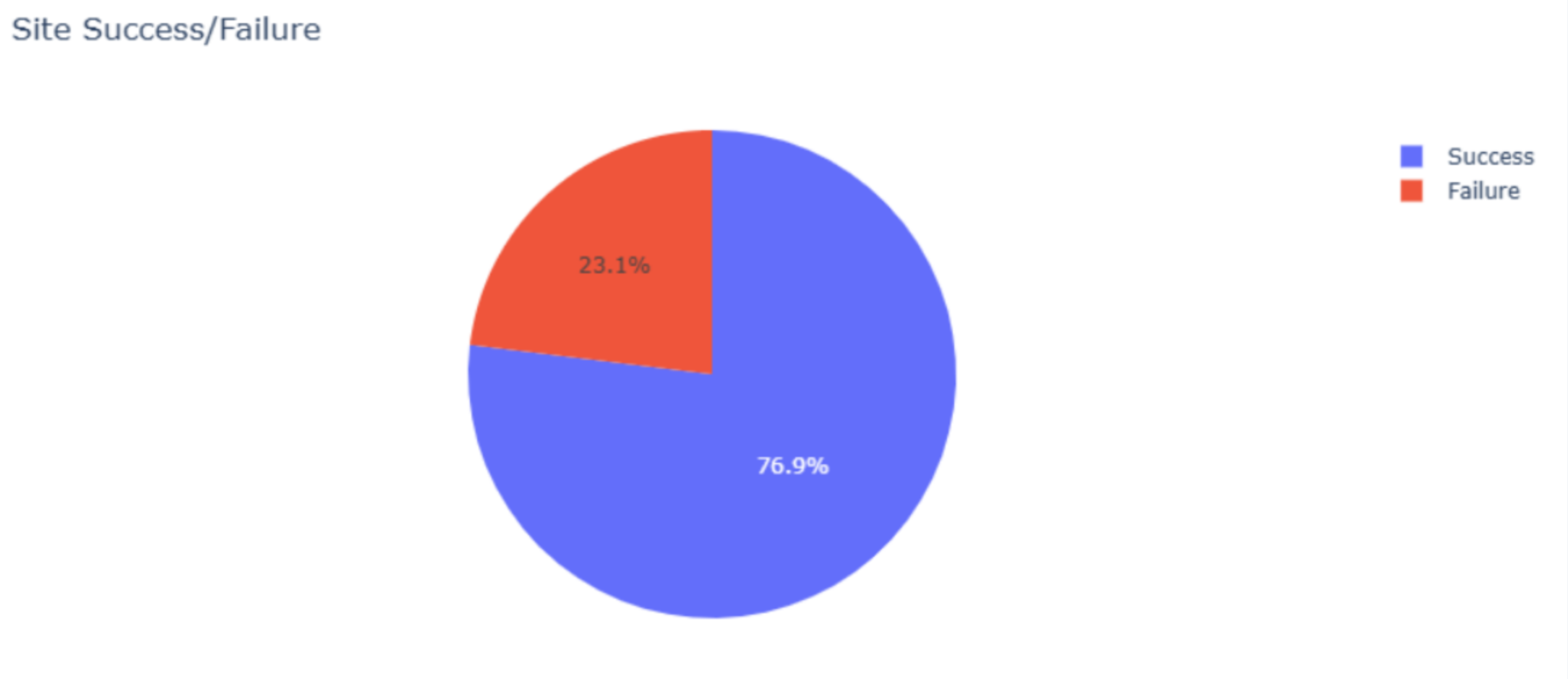

Section 4
Build a Dashboard with Plotly Dash

Landing Success Rate by Launch Site



- KSC LC 39A has the highest landing success rate followed by CCAFS LC-40

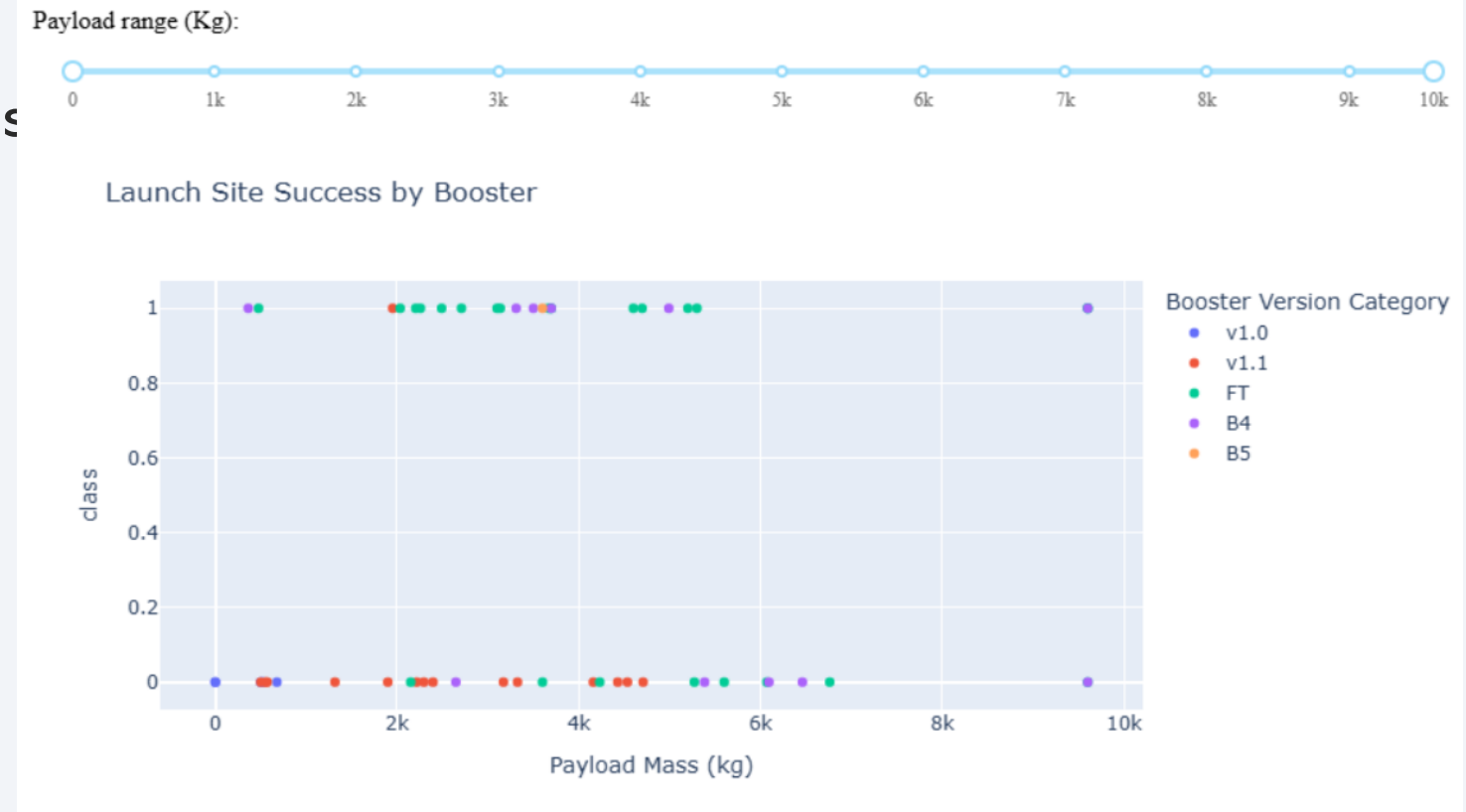
KSC LC-39A Landing Success Rate



- More than 75% of launches result in successful booster landing

Landing Success Rate by Payload and Booster Category

- The FT Booster appears to result in successful landings across the full range of payloads
- There appear to be a high failure rate with Booster v1.0 and v1.1

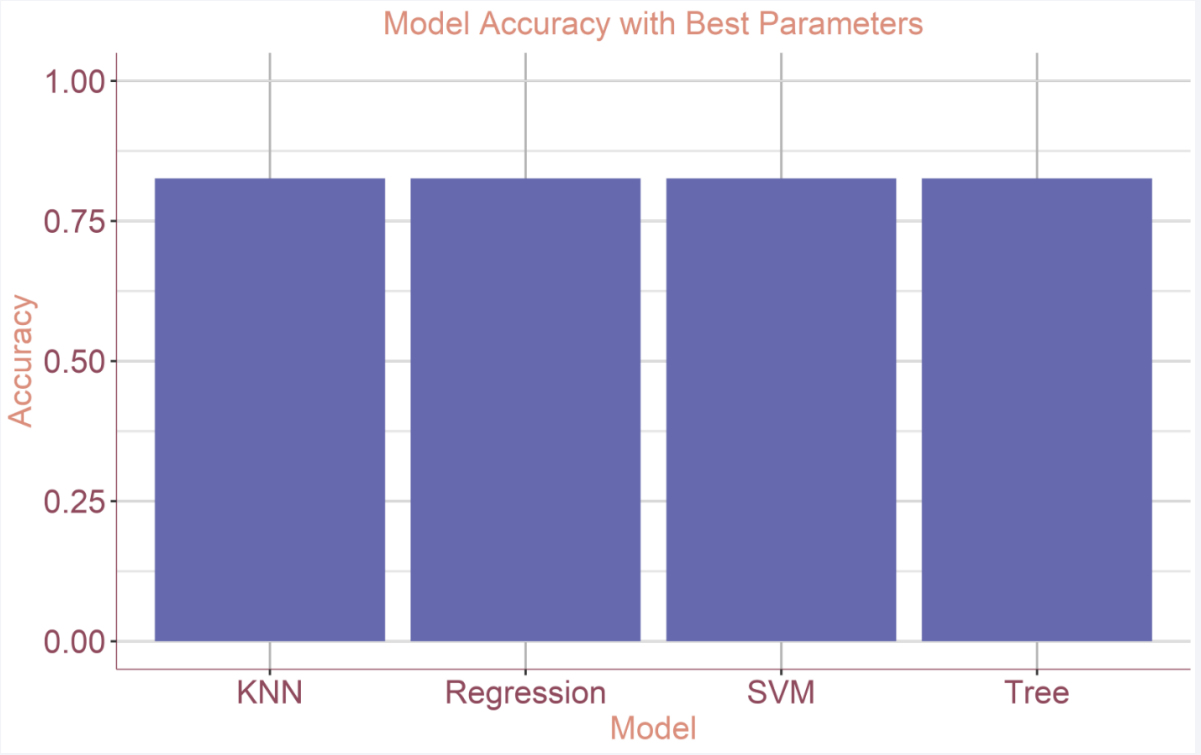


Section 5

Predictive Analysis (Classification)

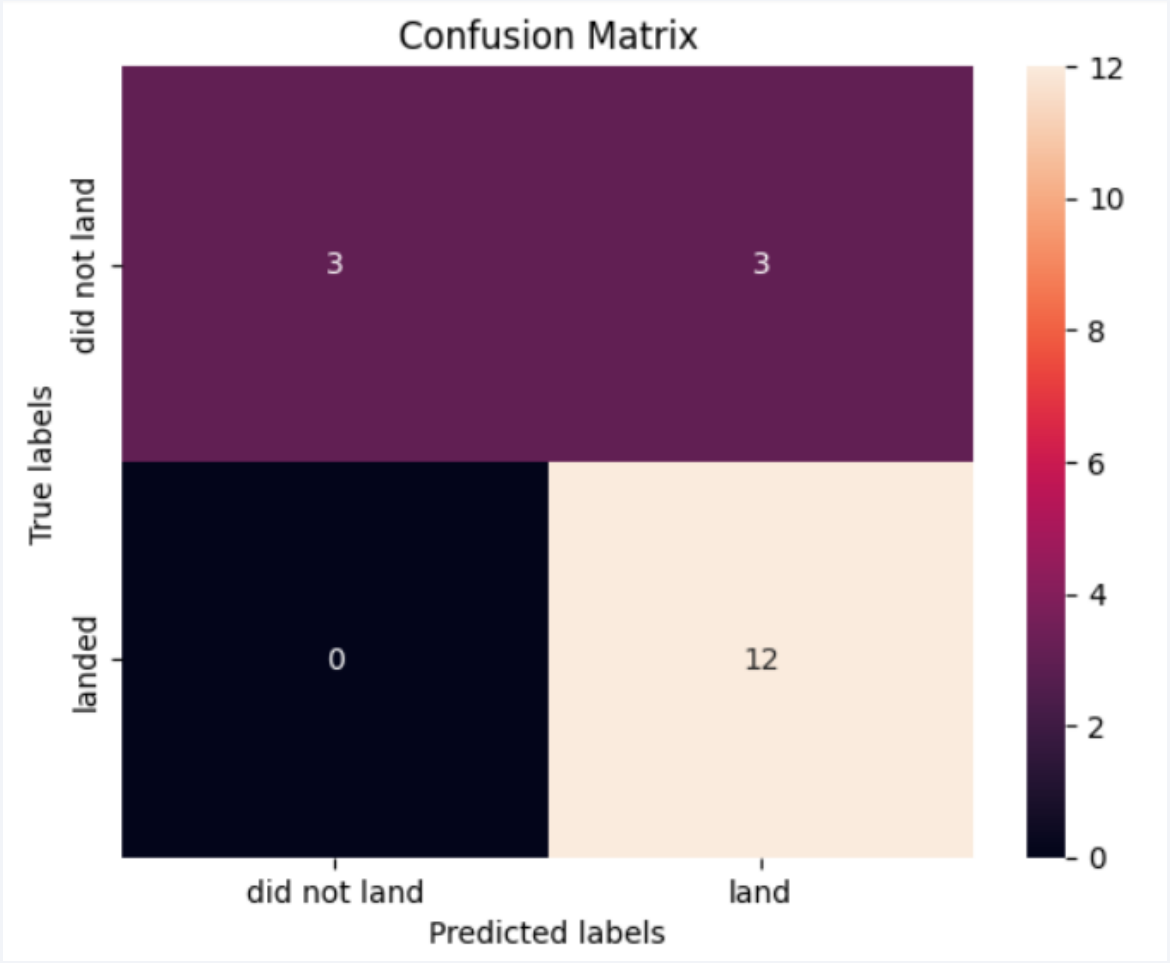
Classification Accuracy

- All models with optimised parameters report identical accuracy.
- This is a surprising finding
  - Triple checked code for all models



Confusion Matrix

- There are 3 false positives
- There are 12 true positives
- False positives appear to be a weakness in the model



Conclusions

- It is possible to reliably predict landing success
- Experience both in terms of the number of flights and number of flights on a booster version improve landing outcomes
- Additional data such as weather metrics may be able to improve the accuracy of the prediction models.

Appendix

- Prediction model tuning and assessment (python code sample)

```
parameters ={'C':[0.01,0.1,1], 'penalty':['l2'], 'solver':['lbfgs']}
model_lr = LogisticRegression()
grid_search = GridSearchCV(estimator=model_lr, param_grid=parameters, cv=10)
logreg_cv = grid_search.fit(X, Y)
best_parameters = grid_search.best_params_
best_model = grid_search.best_estimator_

logreg_cv.fit(X_train, Y_train)
accuracy = logreg_cv.score(X_test, Y_test)
print(f"Accuracy: {accuracy}")

best_model.fit(X_train, Y_train)
accuracy = best_model.score(X_test, Y_test)
print(f"Accuracy: {accuracy}")
```

Thank you!